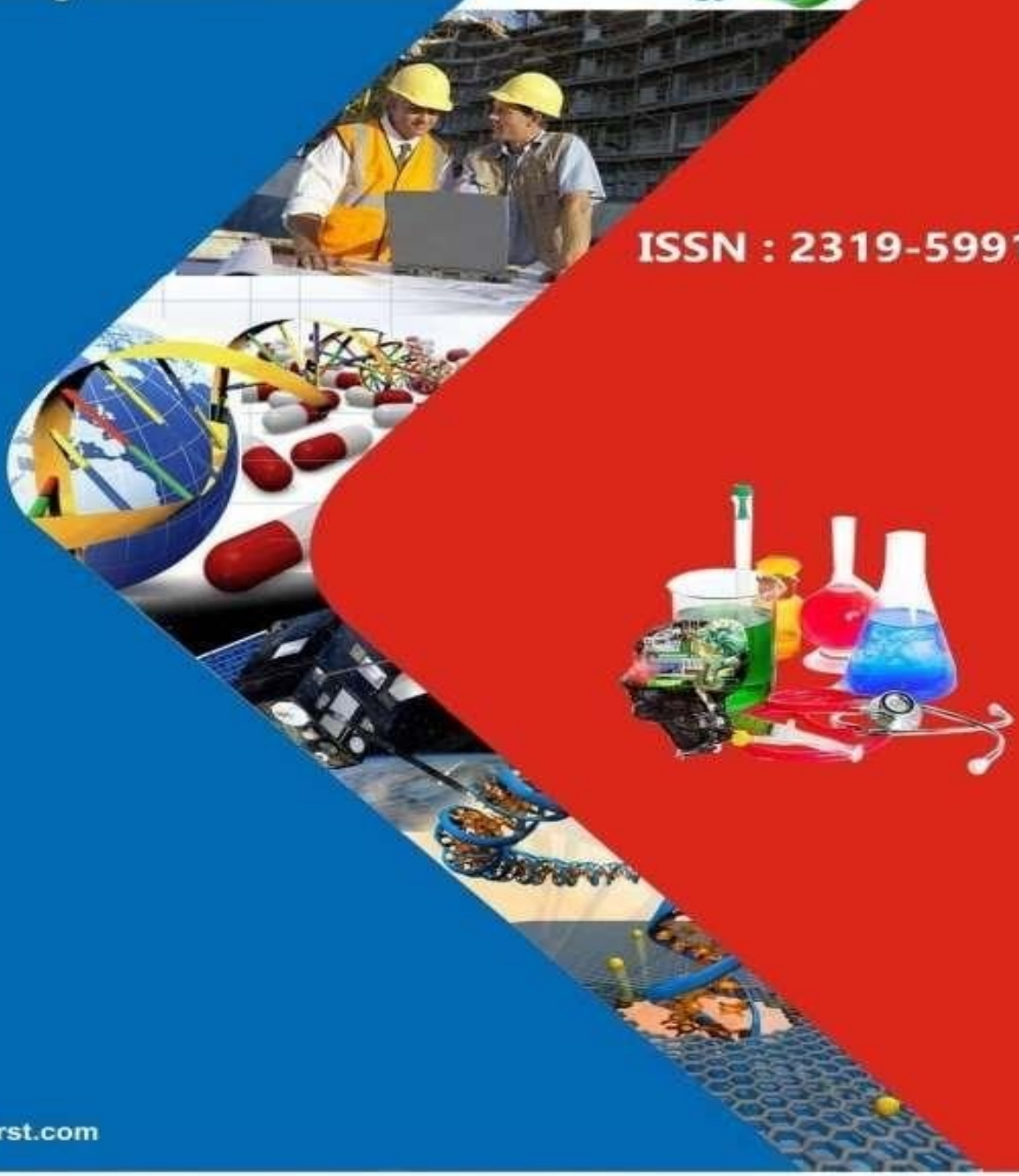
# International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336 Vol. 21, Issue 2, 2025

# DATAFITS: LEVERAGING HETEROGENEOUS DATA FUSION FOR ACCURATE TRAFFIC AND INCIDENT FORECASTING

[1]P.Pavan Kalyan Reddy, MCA Student, Department of MCA

[2]M G K Priyanka,MCA, (Ph.D),Assistant Professor, Department of MCA

[12]Dr KV Subba Reddy Institute of Technology, Dupadu, Kurnool

**ABSTRACT**

In order to create a complete dataset, this study presents DataFITS (Data Fusion on Intelligent Transportation System), an open-source framework that gathers and combines traffic-related data from several sources. According to our hypothesis, a heterogeneous data fusion framework may improve the quality and breadth of information for traffic models, boosting the effectiveness and dependability of applications for Intelligent Transportation Systems (ITS). Two applications that made use of event categorisation and traffic estimate models confirmed our hypothesis. Over the course of nine months, DataFITS gathered four different kinds of data from seven sources and combined them in a spatiotemporal domain. While incident categorisation utilised the k-nearest neighbours (k-NN) method with Dynamic Time Warping (DTW) and Wasserstein metric as distance measurements, traffic estimation models used polynomial regression and descriptive statistics. According to the results, DataFITS enhanced information quality for up to 40% of all roads via data fusion and dramatically expanded road coverage by 137%. Using a polynomial regression model, traffic estimation obtained an R2 score of 0.91, while incident classification reached 90% accuracy on binary tasks (incident or non-incident) and about 80% accuracy on categorising three distinct event categories (accident, congestion, and non-incident).

## I. INTRODUCTION

The design of contemporary Intelligent Transportation Systems (ITSs), which use models to better comprehend different patterns of the transportation system [1], depends critically on the availability of data. This improves both the mobility and safety of people and products. Since contemporary civilisation depends so largely on dependable and efficient transportation, the relevance of these networks has grown significantly in recent years. Both the number of automobiles registered and the number of passengers carried on public transit have significantly increased in Germany alone, hitting all-time highs of 48.5 million cars (2022) and 12.7 billion people carried (2019, prior to the pandemic) [2], [3]. As a consequence, time delays, pollutants, and fuel consumption rise in metropolitan areas, along with the frequency of traffic-related events (such as accidents and congestion) [4].

Because of this, research and industry have been working to develop the next generation of transport systems, which will be economical, environmentally benign, and driven by communication and data analysis technologies. According to our

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336
Vol. 21, Issue 2, 2025

hypothesis, a heterogeneous data fusion framework may improve the quality and coverage of data used as input for traffic models, boosting the effectiveness and dependability of ITS systems. In order to train models for two ITS applications—traffic estimates and incident classification—we therefore suggest the Data Fusion on Intelligent Transportation System (Data FITS) framework, which offers a spatiotemporal fusion of data. Real heterogeneous data (such as weather, traffic, and incidents) is gathered and combined by Data FITS from a variety of sources (such as open databases and map apps). It is then prepared by correcting mistakes, modifying the data structure, and fusing it at the precise place and moment. Data characterisation to measure the advantages of merging diverse data sources and the suggestion of two ITS applications are used to validate our theory. When predicting traffic and categorising events, the two systems' performance validates the advantages of greater data coverage and quality. Therefore, this investigation's primary contributions are:

• A public code repository contains the open-source Data FITS framework for heterogeneous spatiotemporal data fusion, which covers data gathering, processing, and fusion. One

• The description of a heterogeneous dataset that includes actual traffic data from two German cities that was gathered over a nine-month period from seven sources and supplied with the repository.

• A comparison between single and fused datasets; two traffic estimate models, one using descriptive statistics and the other utilising polynomial regression with various factors, including time, road type, and weather.

• Using k-nearest neighbours (k-NN) and Wasserstein and Dynamic Time Warping (DTW) as distance techniques, an event classification model was trained and assessed on heterogeneous fused data.

This is how the remainder of the paper is structured. Section II contrasts our method with contemporary literature that uses data fusion to create applications such as event categorisation and traffic prediction. Section III describes the traffic data applications and the Data FITS architecture. Using the heterogeneous fused data, Section IV assesses our framework's performance as well as the efficacy of our traffic estimating and event categorisation models, confirming our hypothesis. In Section V, we wrap up this work by outlining unresolved issues for further research.

## II. LITERATURE SURVEY

"A survey on big data analytics in intelligent transport systems,"

Y. Wang, B. Ning, T. Tang, F. R. Yu, and L. Zhu

Intelligent transport systems (ITS) research is increasingly focussing on big data, as seen by several global initiatives. A significant quantity of data will be generated by intelligent transportation systems. The large data generated will have a significant influence on how intelligent transportation systems are designed and implemented, making them safer, more effective, and more lucrative. The subject of big data analytics research in ITS is booming. The history and features of big data and intelligent

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

transportation systems are first reviewed in this study. The following section discusses the framework for doing big data analytics in ITS, which includes a summary of the data source and collecting techniques, data analytics platforms and methodologies, and big data analytics application types. The article introduces a number of case studies of big data analytics applications in intelligent transportation systems, such as asset maintenance, public transportation service planning, personal travel route planning, road traffic accident analysis, and rail transportation management and control. Lastly, several unresolved issues with applying big data analytics to ITS are covered in this work.

"Data sharing and real-time processing platform for smart cities,"

S. Sargento, P. Rito, and G. Vitor,

The idea of a smart city necessitates the support of a data platform that can collect, process, and export data from millions of sensors from various sources in a scalable manner for the visualisation, processing, and actuation of historical and real-time data in the city. In order to collect, analyse, visualise, and act upon mobility, environmental, and network data, this article suggests a data platform for the Aveiro Tech City Living Lab. Through a safe and open data platform, the platform's design offers a genuine open platform that third parties may use to gather data and test their own solutions. The findings about the volume of data collected and data samples demonstrate how this platform may be utilised to create new apps and use historical and real-time data for future forecasts and actions in a smart city.

"On the design of virtual sensors for vehicles,"

M. Do Val Machado, A. A. F. Loureiro, P. H. L. Rettore, and A. B. Campolina

Control systems, particularly those for vehicles, depend heavily on physical sensors. In order to maintain a vehicle stable and operational, sensor readings assist drivers in controlling both the vehicle and its internal systems. Not all of the hundreds of precise and varied sensors that are now installed in a contemporary luxury automobile are visible to the driver. Nevertheless, there are certain characteristics and events for which there are no physical sensors. In order to detect complicated variables and ultimately replace and monitor malfunctioning physical sensors, virtual sensors aggregate signals from several sensors to create their own output values depending on circumstances and models. Because of the intricacy of the several processing steps it involves, designing a virtual sensor is often a challenging procedure. The process of developing and prototyping virtual sensors for vehicles is examined in this paper, which also provides examples of virtual sensors made using a framework designed to speed up the design process.

"City data hub: Establishing an interoperable smart city data platform based on standards,"

J. Kim, S. Jeong, and S. Kim,

Similar to the Internet of Things (IoT), smart cities have proliferated in our daily lives. One of the most widely accepted definitions of smart cities is that they tackle urban issues to improve the quality of life for its residents and create sustainable urban

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

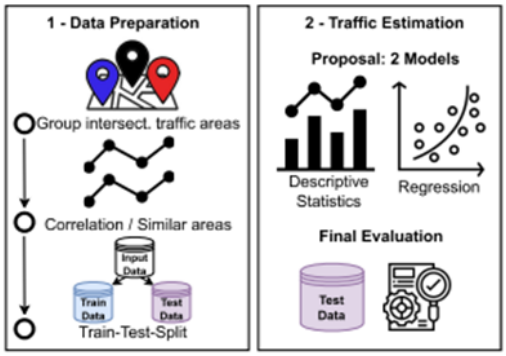
environments. We believe that this may be accomplished by gathering and evaluating data to provide insights from the standpoint of information and communication technologies (ICT). A few instances of problem-solving have been shown, along with the City Data Hub, a standard-based city data platform that has been established. The essential components of smart city platforms have been selected, incorporated into the fundamental architectural principles, and put into practice as a platform. It has been shown that the keys to ensuring ecosystem extensibility and enhancing interoperability are common data models with data markets and standard application programming interfaces (APIs).

"Heterogeneous data fusion from multiple sources,"

L. Zhang, X. Zhang, Y. Xie, and L. Xidao,

The big data age is coming as a result of the internet's exponential rise in data. Big data fusion is a research hotspot because it produces enormous values. But in the age of big data, data exhibits characteristics of high volume, velocity, veracity, and most importantly, diversity, also known as heterogeneity. Data heterogeneity arises from a variety of data sources. Heterogeneous data from several sources presents both possibilities and difficulties for big data fusion. Big data fusion and heterogeneous data fusion techniques are introduced in this study, with a particular emphasis on the use of deep learning techniques in multi-source heterogeneous data fusion. The difficulties of handling heterogeneous data fusion from several sources are also covered.

## SYSTEM ARCHITECTURE




## III. EXISTING SYSTEM

Significant data from actual or virtual sensors is needed to construct ITS applications [5]. A framework for gathering, processing, and exporting heterogeneous data from smart city sensors is presented by Vitor et al. [4], who also provide a variety of statistics and visualisations. Their platform, however, focusses on data security. A smart city data portal with data from several cities is also suggested by [6]. Unlike our framework, we concentrate on enhancing the amount and quality of information by merging data, and we evaluate the benefits of doing so using two ITS applications. By merging data from several sources, data fusion enhances spatiotemporal information [7], [8], [9], and [10]. Data fusion is useful for several applications, including route planning [12] and emergency management [11]. However, further preprocessing is needed to merge different data types and characteristics when fusing heterogeneous data [13], [14]. This study examines two data fusion-supported applications, traffic estimate and incident categorisation, as well

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

as the strategies used to accomplish these objectives, including data collecting, fusion, machine learning, correlation, and various data kinds.

One essential smart city application for improved transport management is traffic estimation. In order to provide precise and trustworthy traffic estimate using historical data, this study focusses on data fusion, spatiotemporal correlation, and machine learning algorithms. Big data presents a chance for heterogeneous data fusion as there is a growing amount of traffic-related data available via open databases (maintained by government agencies) and Application Programming Interfaces (APIs) to commercial apps (Bing, Google Maps, etc.). [15]. Combining data from stationary sensors (like traffic cameras or loop detectors) with information from probing vehicles (such cameras, GPS, cellphone data, or vehicular sensors) is a difficulty. Anand et al. [16] improved a traffic estimate method by fusing journey time (from GPS) and traffic flow data (from cameras) using a Kalman filter. Machine Learning (ML) has been used in several recent traffic estimate models [17], [18], [19], [20], [21], [22], [23], [24], and [25]. Using data from a traffic simulator, Reference [17] suggests an auto-regressive model that adjusts to accidents and other occurrences.

According to their findings, the estimate inaccuracy up to 30 minutes in advance is 12%. In the meanwhile, [18] uses deep learning algorithms to estimate traffic, demonstrating an increase in efficiency and accuracy. These methods talk about using machine learning (ML) to build precise traffic estimate models, however they don't take into account other techniques like data fusion, correlation, etc.

To enhance the quality of traffic estimates, several machine learning techniques use spatiotemporal correlation. A neural network (NN)-based estimating method using Gated Recurrent Unit (GRU) and Graph Convolutional Network (GCN) models is shown in [19] and is openly accessible. The GRU detects dynamic changes in traffic data and records temporal relationships, whereas the GCN records spatial dependencies from the road network. Using data correlation, other NN-based methods, such [20] and [21], demonstrate comparable accuracy gains. An open-source deep learning framework using GCN is proposed by Wang et al. [22] to forecast network-wide traffic many steps in advance. Another open-source approach, the Graph Multi Attention Network (GMAN), is presented by Zheng et al. [23]. It uses an encoder-decoder architecture to enable long-term traffic estimate up to one hour in advance. These methods do not give a way to gather or combine data, but they do use correlation to enhance the models that are being discussed and provide access to their data. Our technique is comparable to the limited literature that uses data fusion, spatiotemporal correlation, and machine learning to estimate traffic. The authors of [26] combine traffic data from both fixed and dynamic sensors while taking into account the spatiotemporal relationship between road segment traffic levels.

The fused data is processed using a Multiple Linear Regression (MLR) model to

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

improve the precision of traffic estimates. In contrast to our solution, this method ignores various data kinds and sources and only uses sensor-derived traffic statistics. To improve traffic estimates, Zhao et al. [24] provide a universal framework for spatiotemporal data fusion. By merging direct and indirect traffic-related data as input for two distinct ML models, the technique presents a fusion strategy to increase accuracy. The indirect traffic-related data characteristics are utilised to enhance the quality of the estimate and include weather and point-of-interest information. While the authors in [24] take into account sites of interest and meteorological conditions, our work concentrates on incident-related data, and their model leverages pre-existing datasets without providing a method for data collecting.

**Disadvantages**

- The data fusion framework Data FITS and data applications traffic estimate and event categorisation were not implemented by the system.
- Since the model does not need it, the fused data from DataFITS is not cleaned and is instead aggregated into traffic zones that comprise one or more road segments.

## IV. PROPOSED SYSTEM

In order to train models for two ITS applications—traffic prediction and incident classification—the system suggests the Data Fusion on Intelligent Transportation System (DataFITS) framework, which offers a spatiotemporal fusion of data. Real heterogeneous data (such as weather, traffic, and incidents) is gathered and combined by DataFITS from a variety of sources (such as open databases and map apps), prepared by correcting mistakes, modifying the data structure, and then fusing them in the precise place and moment. Data characterisation to measure the advantages of merging diverse data sources and the suggestion of two ITS applications are used to validate our theory. When predicting traffic and categorising events, the two systems' performance validates the advantages of greater data coverage and quality.

**Advantages**

• DataFITS is an open-source system for heterogeneous spatiotemporal data fusion that is accessible via a public code repository and covers data gathering, processing, and fusion.

• The description of a heterogeneous dataset that includes actual traffic data from two German cities that was gathered over a nine-month period from seven sources and supplied with the repository.

• A comparison of single and fused datasets; two traffic estimate models, one using descriptive statistics and the other utilising polynomial regression with various factors, including time, road type, and weather.

• Using k-nearest neighbours (k-NN) and Wasserstein and Dynamic Time Warping (DTW) as distance techniques, an event classification model was trained and assessed on heterogeneous fused data.

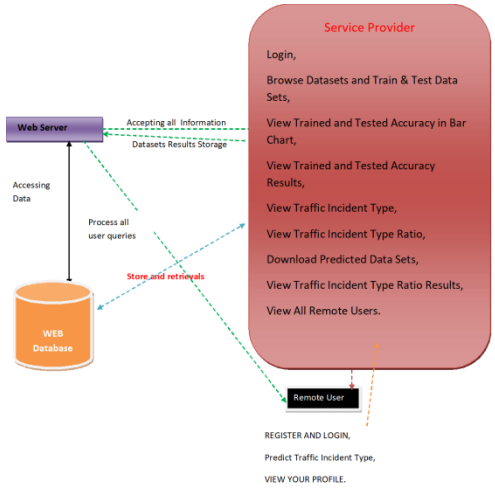
International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

## SYSTEM ARCHITECTURE




## V. IMPLEMENTATION

### Modules description

### Service Provider

The Service Provider must use a working user name and password to log in to this module. He can do many tasks after successfully logging in, including browsing datasets and training and testing datasets. View Traffic Incident Type, View Traffic Incident Type Ratio, Download Predicted Data Sets, Trained and Tested Accuracy in Bar Chart, Trained and Tested Accuracy Results, and View All Remote Users.

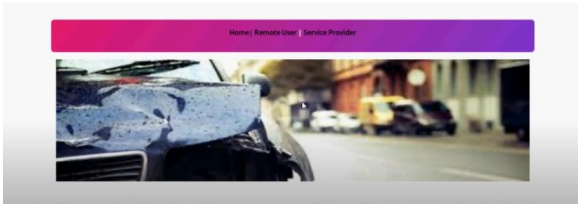
### View and Authorize Users

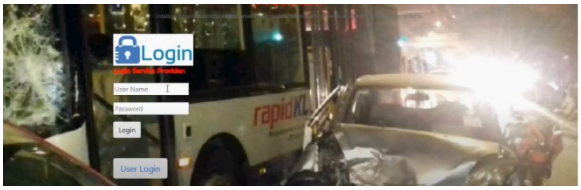
The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

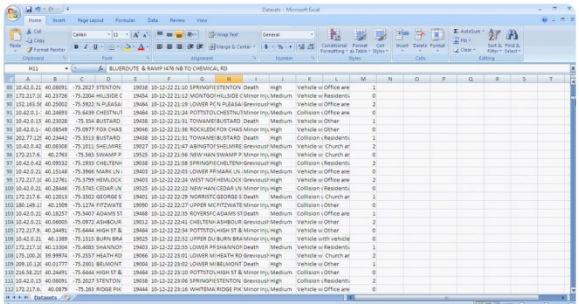
### Remote User

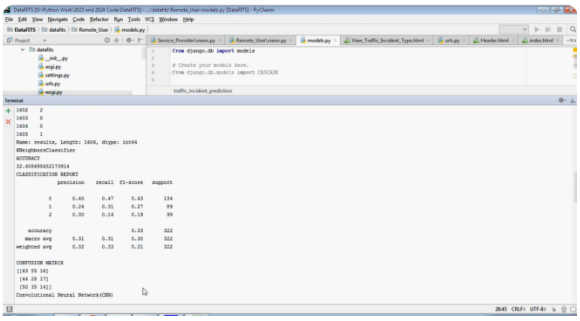
A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. The user will do many tasks after successfully logging in, including registering and logging in, predicting the kind of traffic incident, and seeing their profile.

## VI. SCREENSHOTS



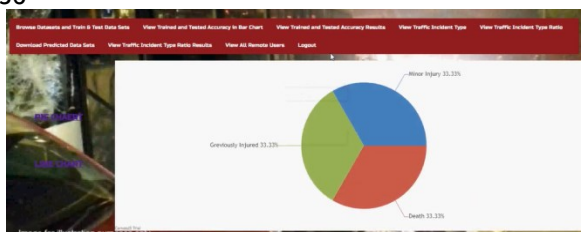


International Journal of Engineering Research and Science & Technology

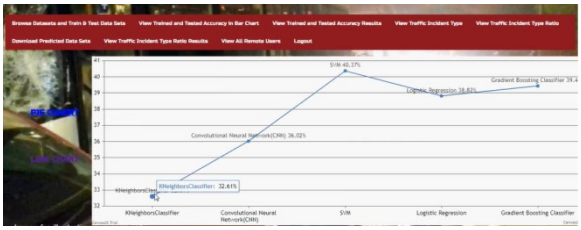
ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

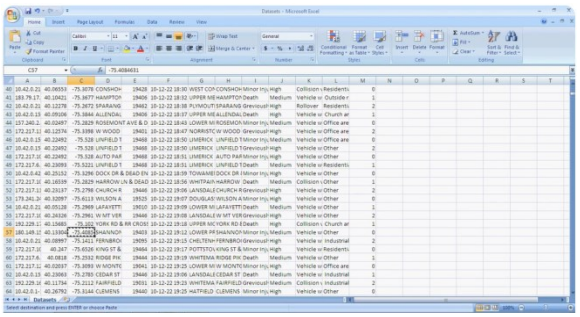
Vol. 21, Issue 2, 2025

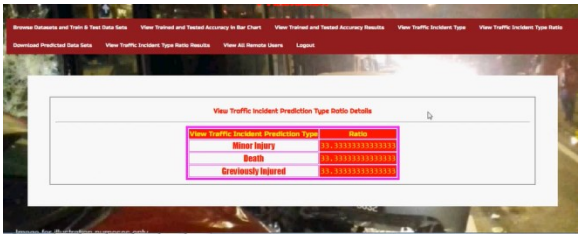





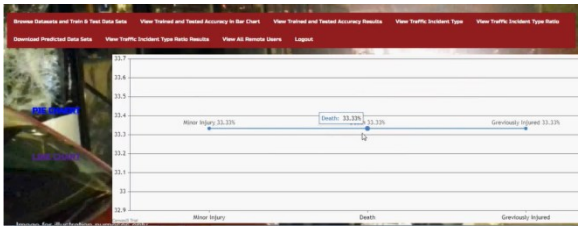



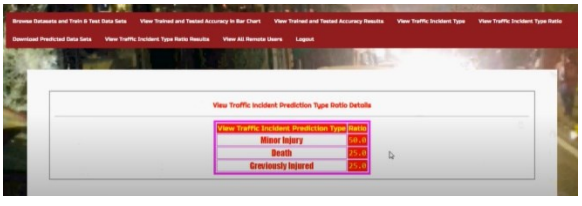








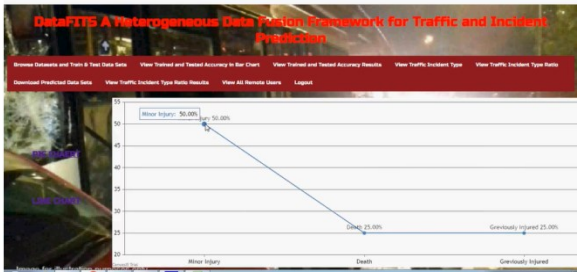



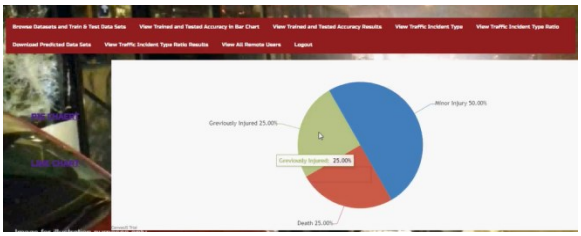
International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336




## VII. CONCLUSION

In this study, we provide Data FITS, an open-source data fusion system that gathers, examines, and combines various types of data. Our hypothesis is that heterogeneous data fusion improves datasets for ITS applications by increasing both the number and quality of data. We created two ITS systems to confirm this: one classified events as accidents, congestion, or non-incidents by combining traffic and incident data, while the other utilised polynomial regression to predict traffic levels. By creating a fused dataset from actual heterogeneous data from two German cities, we were able to quantify the benefits of Data FITS. According to our findings, Data FITS increased the total road coverage by 137% by integrating data from various sources for 40% of all roads. Furthermore, the polynomial regression-based traffic estimating model performed better than our descriptive statistics-based prior method, obtaining a high R2 score of 0.91, low error metrics of 0.05, and reliable traffic predictions using the fused dataset. The fused dataset estimation significantly increased the spatiotemporal coverage of the predicted regions while only slightly improving accuracy as compared to utilising a single sources dataset. By combining traffic and event data, our incident classification model achieves a 90% binary classification accuracy rate in our study. Data preprocessing, such as eliminating ambiguous traffic patterns, increased accuracy by 29% on average. The accuracy of classifying episodes into distinct groups was somewhat lower at 86%, with F1 values indicating that various classes performed differently. We oversampled the training dataset to provide a more consistent representation of the data in order to address this issue, and each class achieved an accuracy of 80%. This issue may also be resolved by gathering further accident data. By gathering and combining more data kinds, enhancing the Data FITS framework's functionality and quality, and broadening its scope of analysis, we want to grow it. We concentrate on data types including photos and social media, which call for techniques like image processing and natural language processing (NLP). Our goal is to compare our present models with alternative models and hyper-parameters for ITS applications using automated machine learning. Additionally, we want to examine the relationship between incidents and traffic and apply it to the models used for traffic estimates. In order to assist strategic operations in urban warfare, we also want to investigate the use of big data in military contexts by merging data from the military and civilian domains. In order to prevent false information from influencing political choices, our system may be improved to

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336

gather and integrate various information kinds (text, image) to provide common operational images and validate information.

**REFERENCES**

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," IEEE Trans. Intell. Transp. Syst., vol. 20, no. 1, pp. 383–398, Jan. 2019.

[2] Umweltbundesamt. (2022). Verkehrsinfrastruktur und fahrzeugbestand. Accessed: Dec. 12, 2022. [Online]. Available:https://www.umweltbundesamt.de/daten/verkehr/verkehrsinfrastrukturfahrzeugbestand

[3] German Federal Statistical Office (Destatis). (2022). Passengers Carried in Germany. Accessed: Jul. 12, 2022. [Online]. Available: https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Transport/Passenger-Transport/Tables/passengerscarried.html

[4] G. Vitor, P. Rito, and S. Sargento, "Smart city data platform for real-time processing and data sharing," in Proc. IEEE Symp. Comput. Commun.(ISCC), Sep. 2021, pp. 1–7.

[5] A. B. Campolina, P. H. L. Rettore, M. Do Val Machado, and A. A. F. Loureiro, "On the design of vehicular virtual sensors," in Proc. 13th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS), Jun. 2017, pp. 134–141.

[6] S. Jeong, S. Kim, and J. Kim, "City data hub: Implementation of standard-based smart city data platform for interoperability," Sensors, vol. 20, no. 23, p. 7000, Dec. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/7000

[7] L. Zhang, Y. Xie, L. Xidao, and X. Zhang, "Multi-source heterogeneous data fusion," in Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD), May 2018, pp. 47–51.

[8] P. H. L. Rettore, B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "Towards intra-vehicular sensor data fusion," in Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC), Nov. 2016, pp. 126–131.

[9] P. H. L. Rettore, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "A method of eco-driving based on intra-vehicular sensor data," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2017, pp. 1122–1127.

[10] P. H. L. Rettore, A. B. Campolina, A. Souza, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Driver authentication in VANETs based on intravehicular sensor data," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jun. 2018, pp. 00078–00083.

[11] G. L. Foresti, M. Farinosi, and M. Vernier, "Situational awareness in smart environments: Socio-mobile and sensor data fusion for emergency response to disasters," J. Ambient Intell. Humanized Comput., vol. 6, no. 2, pp. 239–257, Apr. 2015.

[12] H. Wen, Y. Lin, and J. Wu, "Co-evolutionary optimization algorithm based on the future traffic environment for emergency rescue path planning," IEEE Access, vol. 8, pp. 148125–148135, 2020.

[13] P. H. Rettore, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Vehicular data space: The data point of view," IEEE

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1326-1336
Commun. Surveys Tuts., vol. 21, no. 3, pp. 2392–2418, 3rd Quart., 2019.
[14] S. A. Kashinath et al., “Review of data fusion methods for realtime and multi-sensor traffic flow analysis,” IEEE Access, vol. 9,
pp. 51258–51276, 2021.
[15] W. Jiang and J. Luo, “Big data for traffic estimation and prediction: A survey of data and tools,” Appl. Syst. Innov., vol. 5, no. 1, p. 23, Feb. 2022.